

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant: Junichi TAKEUCHI, et al.
Title: AUTOREGRESSIVE MODEL LEARNING DEVICE FOR
TIME-SERIES DATA AND A DEVICE TO DETECT
OUTLIER AND CHANGE POINT USING THE SAME
Appl. No.: Unassigned
Filing Date: 07/16/2003
Examiner: Unassigned
Art Unit: Unassigned

CLAIM FOR CONVENTION PRIORITY

Commissioner for Patents
PO Box 1450
Alexandria, Virginia 22313-1450

Sir:

The benefit of the filing date of the following prior foreign application filed in the following foreign country is hereby requested, and the right of priority provided in 35 U.S.C. § 119 is hereby claimed.

In support of this claim, filed herewith is a certified copy of said original foreign application:

- Japanese Patent Application No. 2002-207718 filed 07/17/2002.

Respectfully submitted,

Date: July 16, 2003

FOLEY & LARDNER
Customer Number: 22428

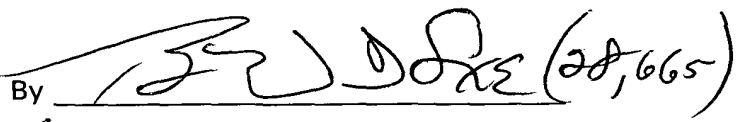



22428

PATENT TRADEMARK OFFICE

Telephone: (202) 672-5407
Facsimile: (202) 672-5399

By


 David A. Blumenthal
Attorney for Applicant
Registration No. 26,257

日 本 国 特 許 庁

JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2002年 7月17日

出 願 番 号

Application Number:

特願2002-207718

[ST.10/C]:

[JP2002-207718]

出 願 人

Applicant(s):

日本電気株式会社

2003年 5月 6日

特 許 庁 長 官
Commissioner,
Japan Patent Office

太田信一郎



出証番号 出証特2003-3032923



【書類名】 特許願

【整理番号】 35001162

【提出日】 平成14年 7月17日

【あて先】 特許庁長官殿

【国際特許分類】 H04L 1/00

【発明者】

 【住所又は居所】 東京都港区芝五丁目7番1号 日本電気株式会社内

 【氏名】 竹内 純一

、 【発明者】

、 【住所又は居所】 東京都港区芝五丁目7番1号 日本電気株式会社内

 【氏名】 山西 健司

【特許出願人】

 【識別番号】 000004237

 【氏名又は名称】 日本電気株式会社

【代理人】

 【識別番号】 100071272

 【弁理士】

 【氏名又は名称】 後藤 洋介

、 【選任した代理人】

、 【識別番号】 100077838

 【弁理士】

 【氏名又は名称】 池田 憲保

【手数料の表示】

 【予納台帳番号】 012416

 【納付金額】 21,000円

【提出物件の目録】

 【物件名】 明細書 1

 【物件名】 図面 1

 【物件名】 要約書 1



【包括委任状番号】 0018587

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 時系列データに対する自己回帰モデル学習装置並びにそれを用いた外れ値および変化点の検出装置

【特許請求の範囲】

【請求項 1】 実数ベクトル値のデータ列を順次読み込みながら該データ列の発生する確率分布を、自己回帰モデルを用いて学習する装置において、自己回帰モデルの十分統計量を、新たに読み込んだデータを用いて過去のデータを忘却しながら更新するデータ更新装置と、該データ更新装置が更新した十分統計量を読み込み、それを用いて自己回帰モデルのパラメータを計算するパラメータ計算装置とを含むことを特徴とする自己回帰モデル学習装置。

【請求項 2】 順次入力される離散値変量と連続値変量との両者または一方で記述されたデータに対してその外れ値スコアおよび変化点スコアを計算して外れ値および変化点を検出する検出装置において、

読み込まれるデータ系列の発生機構を有限個のパラメータで指定される時系列統計モデルとして学習する第一のモデル学習装置と、

該第一のモデル学習装置が学習して得られたパラメータの値を読み込み、読み込んだ時系列モデルのパラメータと入力されたデータとに基づいて各データの外れ値スコアを計算して結果を出力する外れ値スコア計算装置と

を含むことを特徴とする外れ値および変化点の検出装置。

【請求項 3】 請求項 2 において、変化点を検出する検出装置として、

前記外れ値スコア計算装置が計算する外れ値スコアを順次読み込んでその移動平均を計算する移動平均計算装置と、

該移動平均計算装置が計算する外れ値スコアの移動平均を順次読み込んで読み込まれるスコアにおける移動平均の系列の発生機構を有限個のパラメータで指定される時系列統計モデルとして学習する第二のモデル学習装置と、

該第二のモデル学習装置が学習して得られたパラメータの値を読み込み、読み込んだ時系列モデルのパラメータと入力された外れ値スコアの移動平均とに基づいて各移動平均の外れ値スコアを計算しそれをもとのデータの変化点スコアとし

て出力する変化点スコア計算装置と

を更に含むことを特徴とする外れ値および変化点の検出装置。

【請求項4】 請求項3において、前記第一のモデル学習装置が、順次入力されるデータが連続値変量のみで記述される場合、実数ベクトル値のデータ列を順次読み込みながら該データ列の発生する確率分布を、自己回帰モデルを用いて学習するものであって、自己回帰モデルの十分統計量を、新たに読み込んだデータを用いて過去のデータを忘却しながら更新するデータ更新装置と、該データ更新装置が更新した十分統計量を読み込み、それを用いて自己回帰モデルのパラメータを計算するパラメータ計算装置とを含む自己回帰モデル学習装置であることを特徴とする外れ値および変化点の検出装置。

【請求項5】 請求項3において、前記外れ値スコア計算装置および前記変化点スコア計算装置を一つのスコア計算装置とし、離散値変量と連続値変量との両者または一方で記述されたデータの系列に対して、系列中の外れ値および変化点の候補を求める装置として、前記スコア計算装置が計算した外れ値スコアおよび変化点スコアに基づいてデータを降順にソートするソート装置と、該ソート装置がソートした順序に従ってスコアの高いデータを外れ値および変化点の候補として表示する表示装置とを更に含むことを特徴とする外れ値および変化点の検出装置。

【請求項6】 請求項3において、前記外れ値スコア計算装置および前記変化点スコア計算装置を一つのスコア計算装置とし、順次入力される離散値変量と連続値変量の両者または一方で記述されたデータに対して、系列中の外れ値および変化点の候補を求める装置として、前記スコア計算装置が計算した外れ値スコアおよび変化点スコアがあらかじめ定められた閾値を越えたデータを外れ値または変化点の候補として出力するスコア判定装置を更に含むことを特徴とする外れ値および変化点の検出装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、時系列データに対する自己回帰モデル学習装置並びにそれを用いた

外れ値および変化点の検出装置に関し、特に、データ解析技術およびデータマイニング技術に関わり、順次入力される離散値変量と連続値変量との両者または一方で記述されたデータに対してその外れ値スコアおよび変化点スコアを計算して外れ値および変化点を精度よく検出する検出装置に関する。

【0002】

【従来の技術】

従来、この種の時系列データにおける外れ値スコアおよび変化点スコアを計算し外れ値および変化点を検出する検出装置に対しては、統計学、機械学習、データマイニングなどの分野で扱われてきた技術が対象となる。すなわち、本発明で実現する機能である異常値検出と変化点検出とは、従来から統計学、機械学習、データマイニングなどの分野で扱われてきた。

【0003】

しかし、本発明では、データの発生源すなわち情報源に対して定常性を仮定しない状況を対象とする殊になる。

【0004】

こうした場合の外れ値検出については下記のような文献などがある。一つは、バージ (P. Burge) とショーテイラー (J. Shawe-Taylor) とによる方式「Detecting cellular fraud using adaptive prototypes」(Proceedings of AI Approaches to Fraud Detection and Risk Management, pp:9-13, 1997) である。他の一つは、ヤマニシ (K. Yamanishi) ほかによる方式「Online Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms」(Proc. of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, pp:320-324, 2000) である。また他の一つは、ムラド (U. Murad) とピンカス (G. Pinkas

) による方式「Unsupervised profiling for identifying superimposed fraud」(Proceedings of 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, pp:251-261, 1999) である。これらでは、非定常性を扱うために、適応的外れ値検出アルゴリズムが用いられている。

【0005】

また、統計学における通常の変化点検出の方法では、与えられたデータ中の変化点の数をあらかじめ決めておいた上で、変化点間のデータは定常なモデルで記述出来るとして、モデルのあてはめを行うものが知られている。こうした方式については、例えば、下記文献などに記載されている。すなわち、「Journal of American Statistical Association」(69:945-947, 1974) に掲載のグセリー (B. Guthe ry) による論文「Partition regression」、または書籍「Applied Change Point Problems in Statistics」(Nova Science Publishers, Inc, 1995) に掲載のハスコバ (M. Huskova) による論文「Nonparametric procedures for detecting a change in simple linear regression models」等がある。

【0006】

データマイニングにおける変化点検出は、グラニック (V. Guralnik) とスリバスタバ (J. Srivastava) による方式「Event detection from time series data」(Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, pp:32-42, 1999) に記載されている。

【 0 0 0 7 】

【発明が解決しようとする課題】

上述した従来の文献による方式または装置では、時系列データに対する外れ値および変化点の検出装置として、次のような問題点がある。

【 0 0 0 8 】

上述のバージ（P. B u r g e）とショーテイラー（J. S h a w e e - T a y l o r）とによる方式、ヤマニシ（K. Y a m a n i s h i）ほかによる方式、または、ムラド（U. M u r a d）とピンカス（G. P i n k a s）とによる方式といった従来の機械学習技術による逐次処理可能な外れ値検出手法では、時系列データ向きの統計モデルが用いられていない。従って、時系列的性質を持ったデータの特徴を十分に捉えられないという問題があった。ここで、時系列データ向きの統計モデルとは、異なる時点におけるデータ間の相関を表現できるモデルのことであり、例えば、自己回帰モデルまたはマルコフモデルなどがそれにあたる。

【 0 0 0 9 】

また、グラニク（V. G u r a l n i k）とスリバスタバ（J. S r i v a s t a v a）との論文に記載されている従来の変化点検出手法は基本的にデータを一括して処理するいわゆるバッチ処理で動作するものであり、逐次的に処理することが出来ない。さらに、上述した従来の変化点検出手法は局所的に定常であるという仮定を置いて設計されていたが、現実問題としてこのような仮定は不適切であり、除去されるべきである。

【 0 0 1 0 】

さらに、データマイニング等の応用場面では外れ値および変化点を統一的に扱い、両者を検出できることが望ましいにも関わらず、これまでは両者を統一的に扱う枠組みしか知られていなかった。

【 0 0 1 1 】

本発明の課題は、このような問題点を解決することを目的とし、時系列データの性質を捉えられる統計モデルを用い、しかも非定常データに対応することと、外れ値と変化点とを統一的に扱えることと、かつ処理を逐次的に実行できること

とにある。

【0012】

【課題を解決するための手段】

本発明による時系列データに対する外れ値および変化点の検出装置には、実数ベクトル値のデータ列を順次読み込みながら該データ列の発生する確率分布を、自己回帰モデルを用いて学習する装置として、次のような自己回帰モデル学習装置を用いることができる。すなわち、自己回帰モデルの十分統計量を、新たに読み込んだデータを用いて過去のデータを忘却しながら更新するデータ更新装置と、該データ更新装置が更新した十分統計量を読み込み、それを用いて自己回帰モデルのパラメータを計算するパラメータ計算装置とを含むものである。

【0013】

このように、過去のデータを忘却しながら更新する処理するので、時系列データを処理するのに適すると共にその精度を上げることができる。

【0014】

また、本発明による具体的な検出装置は、順次入力される離散値変量と連続値変量との両者または一方で記述されたデータに対してその外れ値スコアおよび変化点スコアを計算して外れ値および変化点を検出する検出装置であって、第一のモデル学習装置と外れ値スコア計算装置と移動平均計算装置と第二のモデル学習装置と変化点スコア計算装置とにより構成されている。

【0015】

第一のモデル学習装置は読み込まれるデータ系列の発生機構を有限個のパラメータで指定される時系列統計モデルとして学習する。外れ値スコア計算装置は、第一のモデル学習装置が学習して得られたパラメータの値を読み込み、読み込んだ時系列モデルのパラメータと入力されたデータとに基づいて各データの外れ値スコアを計算して結果を出力する。移動平均計算装置は、外れ値スコア計算装置が計算する外れ値スコアを順次読み込んでその移動平均を計算する。第二のモデル学習装置は移動平均計算装置が計算する外れ値スコアの移動平均を順次読み込んで読み込まれるスコアにおける移動平均の系列の発生機構を有限個のパラメータで指定される時系列統計モデルとして学習する。変化点スコア計算装置は、第

二のモデル学習装置が学習して得られたパラメータの値を読み込み、読み込んだ時系列モデルのパラメータと入力された外れ値スコアの移動平均とに基づいて各移動平均の外れ値スコアを計算しそれをもとのデータの変化点スコアとして出力する。

【0016】

また、上記第一のモデル学習装置として上述した自己回帰モデル学習装置を用いることができる。

【0017】

また、本発明によると、外れ値スコア計算装置および変化点スコア計算装置を一つのスコア計算装置とし、更に、離散値変数と連続値変数との両者または一方で記述されたデータの系列に対して、系列中の外れ値および変化点の候補を求める装置として、上記スコア計算装置が計算した外れ値スコアおよび変化点スコアに基づいてデータを降順にソートするソート装置と、ソート装置がソートした順序に従ってスコアの高いデータを外れ値および変化点の候補として表示する表示装置とを含むことができる。

【0018】

また、本発明によると、上記ソート装置に代わり、順次入力される離散値変数と連続値変数の両者または一方で記述されたデータに対して、系列中の外れ値および変化点の候補を求める装置として、スコア計算装置が計算した外れ値スコアおよび変化点スコアがあらかじめ定められた閾値を越えたデータを外れ値または変化点の候補として出力するスコア判定装置を含むことができる。

【0019】

【発明の実施の形態】

次に、本発明の実施の形態について図面を参照して説明する。

【0020】

はじめに、記法について説明する。まず「 x 」は実数を成分とする n 次元ベクトル値のデータを表す。また「 y 」は離散値を成分とする m 次元ベクトル値のデータを表す。また「 x 」と「 y 」とをまとめて「 $z = (x, y)$ 」と表す。また、 N 個のデータからなる系列を「 $z^N = z_1 z_2 \cdots z_N$ 」と表す。

【0021】

ここで、このような系列に対して「外れ値スコア」を計算する方法について説明する。

【0022】

まず、データ z を生成する統計モデル

$$p(z_i | \theta) = p(x_i, y_i | \theta)$$

を考える。これは「 z 」の動く範囲、すなわち、レンジ Z 上で定義された確率密度関数のことである。

【0023】

「 θ 」は確率密度を指定するパラメータであり、一般には離散パラメータと連続値パラメータとからなる。こうした確率密度関数としては、例えば「 z 」が連続値変数からなる場合、有限混合ガウス分布、または時系列モデルである自己回帰モデルが用いられる。時系列モデルの場合、 i 番目のデータ z_i の確率密度はそれまでの系列 z^{i-1} に依存するので

$$p(z_i | z^{i-1}, \theta)$$

となる。

【0024】

一般に、外れ値スコアを計算するためには、まずデータ系列に基づいてパラメータ「 θ 」の値を推定する（学習するともいう）。ここで、データの系列を逐次読み込みながら読み込んだデータに基づいて逐次的にパラメータを変更するという「逐次型学習方式」を用いてパラメータを学習する。ここで、データ z_i までを読みこんで学習した結果得られたパラメータの値を「 $\theta^{(i)}$ 」とする。これを用いて「 z_{i+1} 」の外れ値スコアを計算することができる。例えば、対数スコア s_L およびヘリンガースコア s_H は下記数式1および数式2それぞれにより計算される。

【0025】

【数1】

$$s_L = -\log p(z_{i+1} | \theta^{(i)}) \quad (1)$$

【数2】

$$s_H = d^2(p(\cdot|z^1, \theta^{(1)}), p(\cdot|z^{1-1}, \theta^{(1-1)})) \quad (2)$$

ただし「 d^2 」は二つの確率密度間の二乗ヘリンガー距離 (H e l l i n g e r 距離) であり、下記数式3と定義される。

【0026】

【数3】

$$d^2(p, q) = \sum_x \int (\sqrt{p(x, y)} - \sqrt{q(x, y)})^2 dy \quad (3)$$

【0027】

次に、本発明で用いるARモデルについて説明する。ARモデルはn次元実数値ベクトルデータ x_i の系列の確率分布を記述する時系列統計モデルである。まず、補助的な確率変数として「 $\omega^N = \omega_1 \omega_2 \cdots \omega_N$ 」なる系列を導入する。これは「 x 」と同じくn次元であるとする。一般にk次のARモデルは下記数式4で表される。

【0028】

【数4】

$$w_t = \sum \Lambda_i w_{t-i} + \varepsilon \quad (4)$$

で表される。ただし「 A_i 」($i=1, \cdots, k$)はn元正方行列であり、「 ε 」は平均値「0」となる共分散行列 Σ の正規分布に従う確率変数である。

【0029】

いま「 x_i 」が「 u_i 」を用い「 $x_i = u_i + \mu$ 」で与えられるとする。ここで下記数式5が与えられる場合には「 x_t 」の確率密度関数は下記数式6で与えられる。

【0030】

【数5】

$$x_{t-k}^{t-1} = (x_{t-1} \cdots x_{t-k}) \quad (5)$$

【数 6】

$$p(x_t | x_{t-k}^{t-1} : \theta)$$

$$= \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left(- \frac{(x_t - \xi)^T}{2} \Sigma^{-1} (x_t - \xi) \right) \quad (6)$$

ただし、 $\xi = \sum_{i=1}^k \Lambda_i w_{t-i} + \mu$, $\theta = (\Lambda_1, \dots, \Lambda_k, \mu, \Sigma)$

【0031】

外れ値度計算装置は、データの系列を先頭から順に読みこみながら、 i 番目のデータ z_i を読みこんだ際に、その外れ値度 $s(i)$ を出力する。

【0032】

次に、図1を参照して第一の実施例として、上述した自己回帰モデル学習装置について説明する。ここでは、忘却の速さを表す定数 r とARモデルの次数 k があらかじめ与えられているとする。定数 r は「0」から「1」までの間の数値であり、小さいほど、過去のデータを速く忘却することを意味する。

【0033】

図示されるように、第一の実施例は、上述したデータ更新装置で、入力 x_t を受ける忘却型十分統計量計算装置11と、この出力を受けてパラメータ値を送出するパラメータ計算装置12とにより構成される。

【0034】

忘却型十分統計量計算装置11はARモデルにおける忘却型十分統計量を計算する装置である。忘却型十分統計量とは、古いデータの影響ほど小さくなるように修正した十分統計量のことである。ここでいう十分統計量は、 n 次元のベクトル μ と「 $k+1$ 」個の n 元正方行列 C_j ($j=0, 1, \dots, k$) である。忘却型十分統計量計算装置11は、 k 次のARモデルの場合には、 k 時点過去のデータを記憶する機能を有する。

【0035】

パラメータ計算装置12では、与えられた十分統計量に基づいてARモデルのパラメータ $\theta = (\Lambda_1, \dots, \Lambda_k, \mu, \Sigma)$ の値が計算される。

【 0 0 3 6 】

第一の実施例では、以下のように動作する。

【 0 0 3 7 】

まず、データが読み込まれる前にパラメータ記憶装置 1 2 に格納されている各パラメータの値が初期化される。次に、 t 番目のデータが入力される度に、以下のように動作する。

【 0 0 3 8 】

忘却型十分統計量計算装置 1 1 は、データ x_t が入力された際に、これまでに記憶していたデータのうち最も古いものを消去し、代わりに最新のデータ x_t を記憶してデータの列「 $x_t, x_{t-1}, \dots, x_{t-k+1}$ 」を得る。忘却型十分統計量計算装置 1 1 は、これを用いて、保持している十分統計量「 μ, C_j ($j = 0, \dots, k$)」を下記数式 7 および数式 8 に示される更新ルールにより更新し、得られた十分統計量をパラメータ計算装置 1 2 に送りこむ。

【 0 0 3 9 】

【数 7】

$$\mu := (1-r) \mu + r x_t \quad (7)$$

【数 8】

$$C_j := (1-r) C_j + r (x_t - \mu) (x_{t-j} - \mu)^T \quad (8)$$

【 0 0 4 0 】

パラメータ計算装置 1 2 は「 $\bar{\Lambda}_i$ 」($i = 1, \dots, k$) を未知数とする下記数式 9 の連立方程式の解を求める。ただし「 $C_{-j} = C_j$ 」とする。

【 0 0 4 1 】

【数 9】

$$C_j = \sum_{i=1}^k \bar{\Lambda}_i C_{j-i} \quad (j=1, \dots, k) \quad (9)$$

ただし、 $\bar{\Lambda}_i (i=1, \dots, k)$

【0042】

パラメータ計算装置12は、求められた解を「 Λ_i 」に代入し、下記数式10および数式11によりパラメータ θ を計算して、得られたパラメータ θ 「 $= (\Lambda_1, \dots, \Lambda_k, \mu, \Sigma)$ 」を出力する。

【0043】

【数10】

$$x_{it} := \sum_{i=1}^k \Lambda_i (x_{t-1} - \mu) + \mu \quad (10)$$

【数11】

$$\Sigma := (1-r) \Sigma + r (x_t - z_{it}) (x_t - z_{it})^T \quad (11)$$

【0044】

次に、図2を参照して第二および第三の実施例について説明する。

【0045】

図示されるように、この実施例では、上述した第一及び第二のモデル学習装置に対応する時系列モデル学習装置21、24、移動平均計算装置22、および上述した外れ値スコア計算装置および変化点スコア計算装置の両者を包含するスコア計算装置23により構成される。第二の実施例は時系列モデル学習装置21および外れ値スコア計算装置により実現し、第三の実施例は時系列モデル学習装置21、24およびスコア計算装置23により実現される。

【0046】

時系列モデル学習装置21、24は、逐次的に読み込みながら、時系列モデルの確率密度関数におけるパラメータを学習する装置である。

【0047】

ただし、一方の時系列モデル学習装置21は入力されるデータ z_t に関する確率密度関数を学習する装置であり、用いる確率密度関数 F_p は下記数式12で表わされる。

【0048】

【数 12】

$$F_p = p(z_t | z^{t-1}, \xi) \quad (12)$$

【0049】

他方の時系列モデル学習装置 24 は、移動平均計算装置 23 が計算するスコアの移動平均の系列に関する確率密度関数を学習する装置であり、k 次の一変量 AR モデルを用いる。その確率密度関数 F_{qk} は下記数式 13 で表わされる。

【0050】

【数 13】

$$F_{qk} = q(\alpha_t | \alpha_{t-k}^{t-1}, \theta) \quad (13)$$

【0051】

スコア計算装置 23 は、確率密度関数 F_p 、 F_{qk} のパラメータとデータとを読み込み、データ x_t のスコアを計算する。スコア計算装置 23 は、計算機能のほか「 z_t 」の系列については最近 $u(z)$ 個のデータを、「 α_t 」の系列については最近 $u(\alpha)$ 個のデータを、また、「 θ 」と「 ξ 」とのそれぞれについては一つ前のパラメータを保存する機能を有する。例えば、k 次の AR モデルを用いる確率密度関数 F_{qk} の場合、「 $u(\alpha) = k$ 」の条件で、対数スコアまたはヘリンガースコアを計算することができる。

【0052】

移動平均計算装置 22 は、逐次的に入力される実数値のデータに対しその T 移動平均を計算して出力する装置である。そのために移動平均計算装置 22 は、内部に T 個の実数値を格納する機能を有する。

【0053】

第二の実施例に関する装置は以下の順序で動作する。

【0054】

まず、装置全体の初期化が行われ、パラメータおよびデータを記憶する装置においてはあらかじめ定められた適当な値がセットされる。図示される装置は、t 番目のデータ z_t 「 $= (x_t, y_t)$ 」が入力されるたびに、以下のように動作

する。

【0055】

時系列モデル学習装置21およびスコア計算装置23はデータ z_t の入力を受ける。スコア計算装置23は、過去に入力されて保存してある確率密度関数 F_p のパラメータ ξ と、入力されたデータ z_t および過去のデータ「 z_{t-1} , z_{t-2} , \dots , z_{t-u} 」とに基づいてデータ z_t のスコアを外れ値スコア s_t として計算し、得られた外れ値スコア s_t を移動平均計算装置22に送り込むと同時に外部へ出力する。

【0056】

第三の実施例に関する装置は上記第二の実施例に続く以下の順序で動作する。

【0057】

移動平均計算装置22は、スコア計算装置23からスコア s_t を送り込まれると、保存してある最も古いスコアを消去し、新しく入力されたスコア s_t を保存する。次いで、移動平均計算装置22は、保存されているT個のスコアの平均値 α_t を計算して、時系列モデル学習装置24に送り込む。

【0058】

時系列モデル学習装置24は、上記第一の実施例で説明した通りに動作してk次の一変量ARモデルを用いる確率密度関数 F_{qk} のパラメータ ξ を更新し、得られたパラメータ θ とスコア α_t とをスコア計算装置23に送り込む。スコア計算装置23は、過去に入力されて保存してある下記数式14の確率密度関数 F_q のパラメータ θ と、入力されたデータ α_t および過去のデータ「 α_{t-1} , α_{t-2} , \dots , α_{t-u} 」とに基づいてスコア α_t のスコア、すなわち変化点スコアを計算し、得られたスコアを出力する。

【0059】

【数14】

$$F_q = q(\alpha_t | \alpha^{t-1}, \theta) \quad (14)$$

【0060】

次に、図3を参照して第四の実施例について説明する。

【 0 0 6 1 】

ここでは、データ 3 1、上述したスコア計算装置である外れ値スコアおよび変化点スコア計算装置 3 2、スコア付きデータ 3 3、ソート装置 3 4、および表示装置 3 5 が示されている。データ 3 1 は有限の長さのデータ系列を蓄えたデータベースである。外れ値スコアおよび変化点スコア計算装置 3 2 は上記実施例 2 または実施例 3 で説明した外れ値スコアおよび変化点スコアを計算する装置である。スコア付きデータ 3 3 は外れ値スコアおよび変化点スコア計算装置 3 2 の出力を受けて蓄積する。ソート装置 3 4 はデータを外れ値スコアと変化点スコアとを用いてスコアの高い順にソートする。

【 0 0 6 2 】

図示された装置は以下の順序で動作する。外れ値スコアおよび変化点スコア計算装置 3 2 は、データ 3 1 にアクセスしてデータ系列を順に読み込みつつ各データについて外れ値スコアおよび変化点スコアを計算し、スコア付きデータ 3 3 にデータ、外れ値スコア、および変化点スコアの三つ組みを順次送り込む。スコア付きデータ 3 3 は送られてきたデータを蓄積する。ソート装置 3 4 は、スコア付きデータ 3 3 のデータベースにアクセスして格納されているデータを、外れ値スコアと変化点スコアとを用いスコアの高い順にソートして表示装置 3 5 に送り込む。表示装置 3 5 は送られてきた 2 種類のソート済みデータをソートされた順にリストして表示する。

【 0 0 6 3 】

次に、図 4 を参照して第五の実施例について説明する。

【 0 0 6 4 】

ここでは、データ 4 1、上述したスコア計算装置である外れ値スコアおよび変化点スコア計算装置 4 2、スコア付きデータ 4 3、スコア判定装置 4 4、および表示装置 4 5 が示されている。図 4 のスコア判定装置 4 4 が図 3 のソート装置 3 4 の代わりに備えられている。

【 0 0 6 5 】

データ 4 1 は有限の長さのデータ系列を蓄えたデータベースである。外れ値スコアおよび変化点スコア計算装置 4 2 は上記実施例 2 または実施例 3 で説明した

外れ値スコアおよび変化点スコアを計算する装置である。スコア付きデータ 43 は外れ値スコアおよび変化点スコア計算装置 42 からの出力を受けて蓄積する。スコア判定装置 44 は、スコア付きデータ 43 のデータベースにアクセスしてデータを外れ値スコアと変化点スコアとを用いて予め定められた閾値を越えたデータを表示装置 45 に送り込む。

【0066】

図示された装置は以下の順序で動作する。外れ値スコアおよび変化点スコア計算装置 42 は、データ 41 のデータベースにアクセスして、データ系列を順に読み込みつつ、各データについて外れ値スコアおよび変化点スコアを計算する。スコア付きデータ 43 のデータベースに、データ、外れ値スコア、および変化点スコアの三つ組みを順次送り込む。ブロック 43 のデータベースは送られてきたデータを蓄積する。スコア判定装置 44 は、スコア付きデータ 43 のデータベースにアクセスして格納されているデータから、外れ値スコアと変化点スコアとを用いて予め定められた閾値を越えたデータを表示装置 45 に送り込む。表示装置 45 は送られてきた 2 種類のデータをそのまま、またはソートされた順にリストアップして表示する。

【0067】

次に図 5 を参照して、図 2 を参照して説明した外れ値スコアおよび変化点スコアのスコア計算装置を用いて解析した実データについて説明する。

【0068】

この実験は変化点検出を目的に行った。これは、東証株価指数 (TOPIX) の日次データ (1946 年 - 1998 年) を解析した例であり、このうち 1985 年から 1995 年までの結果を示している。グラフには、もとのデータ、およびそれに付けられた変化点スコアを描いてある。データには前処理が行われてある。すなわち、もとの系列を「一次元」の「 x_t 」とするとき、これを「 x_t 、 $x_t - x_{t-1}$ 」と変換してある。これによって、平均的値の変化だけでなく、トレンドの急激な変化も検出できることが期待出来る。この解析結果によると、いわゆるブラックマンデー、またはバブル経済の発生および崩壊などの時期に変化点スコアが高くなっていることが分かる。グラフにおけるブラックマンデーに

ついては、その翌日に著しく高いピークが示されている。

【0069】

上記説明では、図示された機能ブロックを参照しているが、機能の分離併合による配分などの変更は上記機能を満たす限り自由であり、上記説明が本発明を限定するものではない。

【0070】

【発明の効果】

以上説明したように本発明によれば、データの時系列において、その中に現れる統計的な外れ値および変化点であることを示す度合いを外れ値スコアおよび変化点スコアとして計量し、高精度でそれらの検出を行なうことができるという効果を得ることができる。

【0071】

その理由は、まず、順次入力されるデータ列について、読み込まれるデータ系列の発生機構を時系列統計モデルとして学習する時系列モデル学習装置を用いているからである。また、スコア計算装置が、時系列モデルのパラメータと入力されたデータとに基づいて各データの外れ値スコアを計算しているからである。更に、外れ値スコアの移動平均を計算する移動平均計算装置と、移動平均の系列の発生機構を時系列統計モデルとして学習する時系列モデル学習装置と、外れ値スコアの移動平均に基づいて移動平均の外れ値スコアを更に計算してそれをもとのデータの変化点スコアとして出力するスコア計算装置とを組み合わせ、外れ値スコアおよび変化点スコアを計算することにより外れ値および変化点を検出しているからである。

【図面の簡単な説明】

【図1】

本発明によるARモデル学習装置の実施の一形態を示す構成図である。

【図2】

本発明による外れ値スコアおよび変化点スコアの計算装置における実施の一形態を示す構成図である。

【図3】

本発明による外れ値および変化点の候補を求める装置における実施の一形態を示す構成図である。

【図 4】

図 3 とは異なる本発明による外れ値および変化点の候補を求める装置における実施の一形態を示す構成図である。

【図 5】

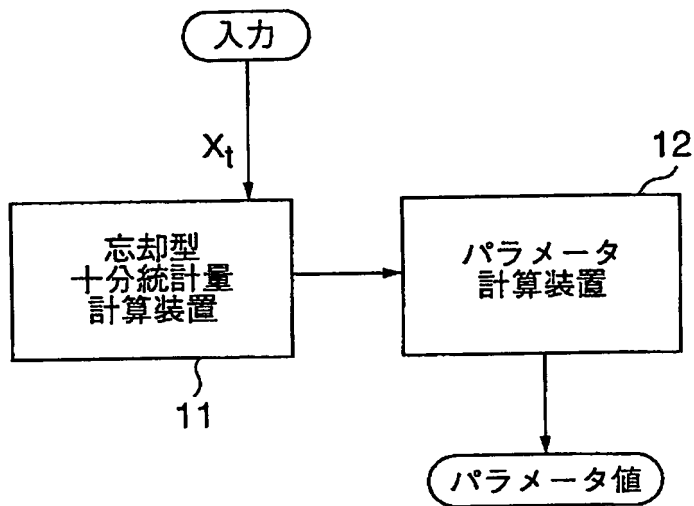
図 2 に示されるスコア計算装置を用いた変化点についての実験結果の一実施例を示すグラフである。

【符号の説明】

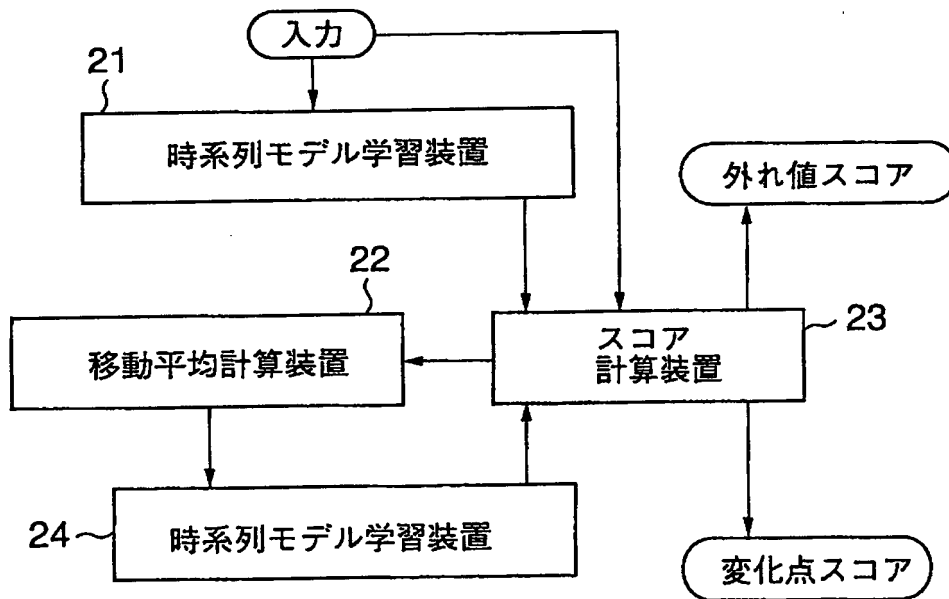
- 1 1 忘却型十分統計量計算装置
- 1 2 パラメータ計算装置
- 2 1、2 4 時系列モデル学習装置
- 2 2 移動平均計算装置
- 2 3 スコア計算装置
- 3 1、4 1 データ（データベース）
- 3 2、4 2 外れ値スコアおよび変化点スコア計算装置
- 3 3、4 3 スコア付きデータ（データベース）
- 3 4 ソート装置
- 3 5、4 5 表示装置
- 4 4 スコア判定装置

【書類名】 図面

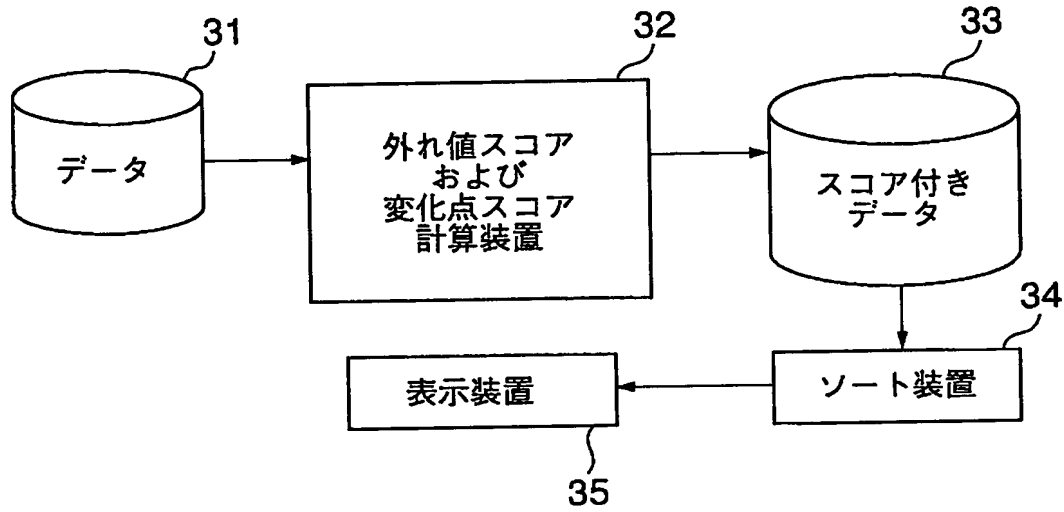
【図 1】



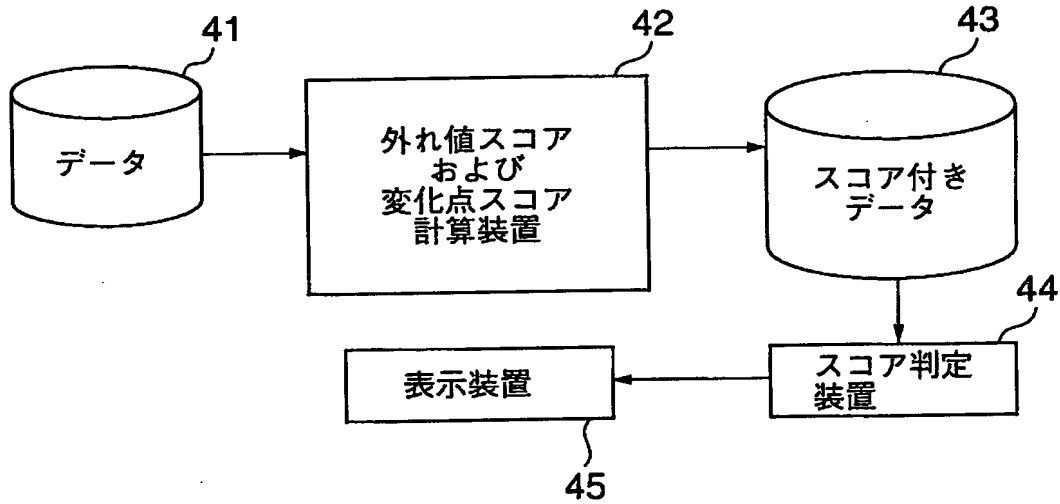
【図 2】



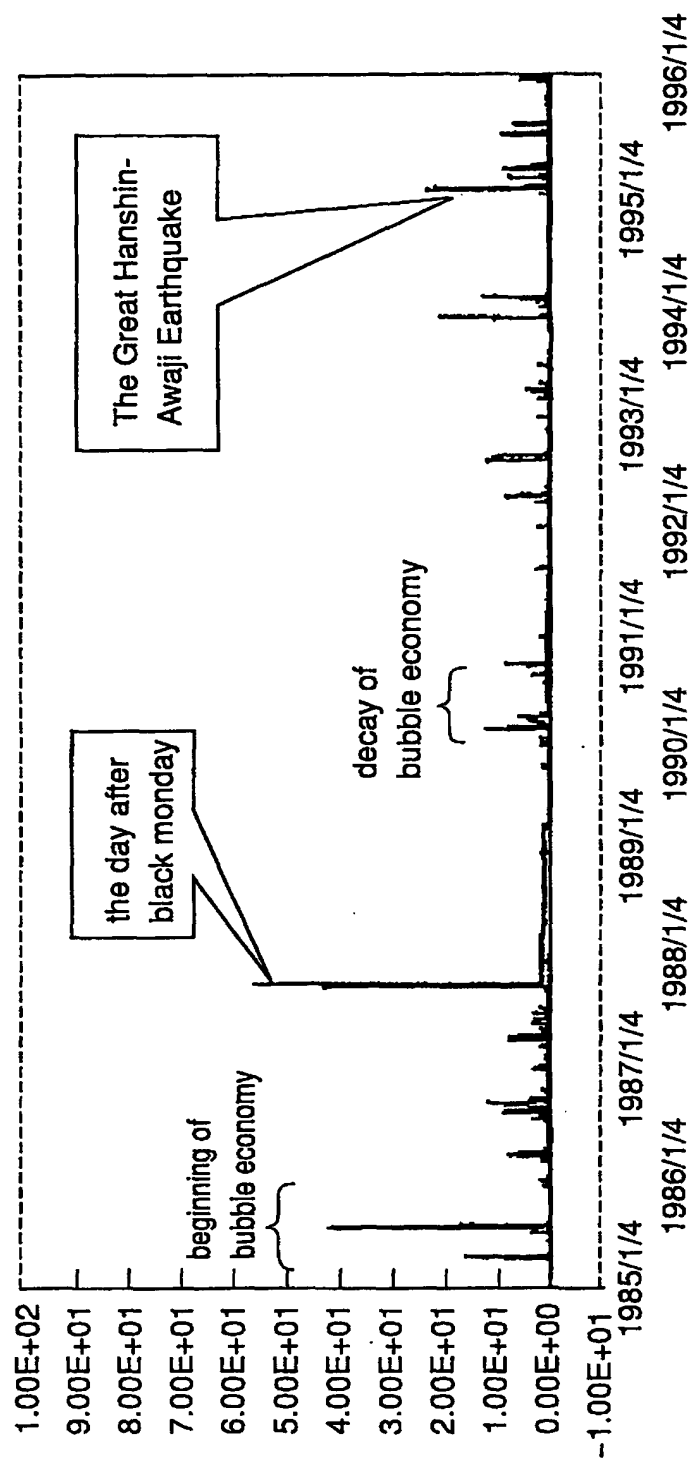
【図3】



【図4】



【図5】



【書類名】 要約書

【要約】

【課題】 データの時系列中に現れる統計的な外れ値および変化点を示す度合いを計量し、外れ値および変化点の高精度な検出を行なうことができる。

【解決手段】 順次入力されるデータ列について、読み込まれるデータ系列の発生機構を時系列統計モデルとして学習する時系列モデル学習装置 21 と、時系列モデルのパラメータと入力されたデータとに基づいて各データの外れ値スコアを計算するスコア計算装置 23 と、外れ値スコアの移動平均を計算する移動平均計算装置 22 と、移動平均の系列の発生機構を時系列統計モデルとして学習する時系列モデル学習装置 24 と、外れ値スコアの移動平均に基づいて移動平均の外れ値スコアを更に計算してそれをもとのデータの変化点スコアとして出力する上記スコア計算装置 23 とを組み合わせ、外れ値スコアおよび変化点スコアを計算することにより外れ値および変化点を検出している。

【選択図】 図 2

出 願 人 履 歴 情 報

識別番号 [000004237]

1. 変更年月日 1990年 8月29日
[変更理由] 新規登録
住 所 東京都港区芝五丁目7番1号
氏 名 日本電気株式会社